A Bayesian Model for Joint Learning of Categories and their Features

A Bayesian Model for Joint Learning of Categories and their Features

Lea Frermann and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
l.frermann@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Categories such as ANIMAL or FURNITURE are acquired at an early age and play an important role in processing, organizing, and conveying world knowledge. Theories of categorization largely agree that categories are characterized by features such as `function` or `appearance` and that feature and category acquisition go hand-in-hand, however previous work has considered these problems in isolation. We present the first model that jointly learns categories and their features. The set of features is shared across categories, and strength of association is inferred in a Bayesian framework. We approximate the learning environment with natural language text which allows us to evaluate performance on a large scale. Compared to highly engineered pattern-based approaches, our model is cognitively motivated, knowledge-lean, and learns categories and features which are perceived by humans as more meaningful.

1 Introduction

Categorization is one of the most basic cognitive functions. It allows individuals to organize their subjective experience of their environment by structuring its contents. This ability to group different objects into the same category based on their common characteristics underlies major cognitive activities such as perception, learning, and the use of language. Global categories (such as FURNITURE or ANIMAL) are shared among members of societies, and influence how we perceive, interact with, and argue about the world.

Given its fundamental importance, categorization is one of the most studied problems in cognitive science. The literature is rife with theoretical and experimental accounts, as well as modeling simulations focusing on the emergence, representation, and learning of categories. Most theories assume that basic level concepts such as *dog* or *chair* are characterized by features such as `barks` or `used-for-sitting`, and are grouped into categories based on those features. Although the precise grouping mechanism has been subject to considerable debate (including arguments in favor of *exemplars* (Nosofsky, 1988), *prototypes* (Reed, 1972), and category *utility* (Corter and Gluck, 1992)), it is fairly uncontroversial that categories are associated with featural representations.

Experimental studies show that the development of categories and feature learning mutually influence each other (Goldstone et al., 2001; Schyns and Rodet, 1997): concepts are categorized based on their features, but the perception of features is influenced by already established categories, and, like categories, features evolve over time. There is also evidence that features such as `barks` or `runs` are grouped into types like `behavior` (Ahn, 1998; McRae et al., 2005; Spalding and Ross, 2000), and the distribution of feature types varies across categories. For instance, living-things such as ANIMALS have characteristic `behavior`, whereas artifacts such as TOOLS have characteristic `functions`, and both categories have characteristic `appearance`.

In this paper, we investigate the problem of *jointly* learning categories and their feature types. Previous modeling work has largely considered these problems in isolation, focusing either on category learning with a fixed set of simplistic features (Anderson, 1991; Sanborn et al., 2006) or feature learning (Austerweil and Griffiths, 2013; Baroni et al., 2010;

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1576–1586,
Denver, Colorado, May 31 – June 5, 2015.

Kelly et al., 2014), but not both.

We present a Bayesian model which induces (semantic) categories and feature types from natural language text. Although language is one of many factors influencing category formation (others include the physical world, how we perceive it, and interact with it), large text corpora encode a surprising amount of extralinguistic information (Riordan and Jones, 2011), and can thus be viewed as an approximation of the learning environment. Moreover, focusing on textual data, allows us to build categorization models with theoretically unlimited scope, and evaluate categories and their features on a much larger scale than previous work in the cognitive science literature.

Our model induces categories (e.g., ANIMALS) and their feature types (e.g., `behavior`) from observations of target concepts (e.g., *lion*, *cow*) and their co-occurring contexts (e.g., *eats*, *sleeps*, *large*). While we can directly evaluate learnt categories through comparison against behavioral data, evaluating feature types is less straightforward. Previous work has shown that the kinds of features learnable from text are qualitatively different from those produced by humans, which makes direct comparison difficult (Baroni et al., 2010; Kelly et al., 2014). We circumvent this problem by assessing in a crowd-sourcing experiment whether the induced feature types are *relevant* for a given category and whether they form a *coherent* class. Evaluation results show that our joint model learns accurate categories and feature types achieving results competitive with highly engineered approaches focusing exclusively on feature learning.

2 Related Work

The problems of category formation and feature learning have been considered largely independently in the literature. Bayesian categorization models were pioneered by Anderson (1991) and recently reformalized by Sanborn et al. (2006). These models are aimed at replicating human behavior in small scale category acquisition studies, where a fixed set of simple (e.g., binary) features is assumed. Frermann and Lapata (2014) propose a model similar in spirit, which they apply to large scale corpora, while investigating incremental learning in the context of child category acquisition (see also Fountain and Lapata (2011) for a non-Bayesian approach). Their model associates sets of features with categories as a by-product of the learning process, however these feature sets are independent across categories and are not optimized during learning.

Previous approaches on feature learning have primarily focused on emulating or complementing norming studies by automatically extracting norm-like properties from textual corpora (e.g., *elephant* `has-trunk`, *scissors* `used-for-cutting`). A common theme in this line of research is the use of pre-defined syntactic patterns (Baroni et al., 2010), or manually created rules specifying possible connection paths of concepts to features in dependency trees (Devereux et al., 2009; Kelly et al., 2014). Once extracted, the features are typically weighted in order to filter out noisy instances. Features are learnt for individual concepts rather than categories. Austerweil and Griffiths (2013) also focus exclusively on feature learning, however from sensory data. They develop a nonparametric Bayesian model which is able to infer unlimited features, based on distributional patterns as well as category information.

To our knowledge, we propose the first Bayesian model that jointly learns categories and their features, arguing that the two tasks are mutually dependent. Our model is knowledge-lean, it learns from raw text in a single process, without relying on parsing resources, manually crafted rule patterns, or post-processing steps. Our work also differs from approaches which combine topic models with human-produced feature norms (Steyvers, 2010). Our aim is not to boost the generalization performance of a topic model, rather we investigate how both categories and features can be jointly learnt from data.

3 The BCF Model

In this section we present our Bayesian model of category and feature induction (henceforth, BCF). BCF jointly learns categories, feature types, and their associations. Specifically, it infers one global set of feature types which is shared across categories (e.g., ANIMALS and VEHICLES can be described in terms of `colors`). However, categories

Generate category distribution, $\theta \sim Dir(\alpha)$
for concept type ℓ **do**
  Generate category, $k^{\ell} \sim Mult(\theta)$
for category k **do**
  Generate feature type distribution, $\mu_k \sim Dir(\beta)$
for feature type g **do**
  Generate feature distribution, $\phi_g \sim Dir(\gamma)$
for stimulus d **do**
  Observe concept c^d and retrieve category k^{c^d}
  Generate a feature type, $g^d \sim Mult(\mu_{k^{c^d}})$
  for feature position i **do**
    Generate a feature $f_{d,i} \sim Mult(\phi_{g^d})$

Figure 1: The generative story of the BCF model. Observations f and latent labels k and g are drawn from Multinomial distributions (*Mult*). Parameters for the multinomial distributions are drawn from Dirichlet distributions (*Dir*).

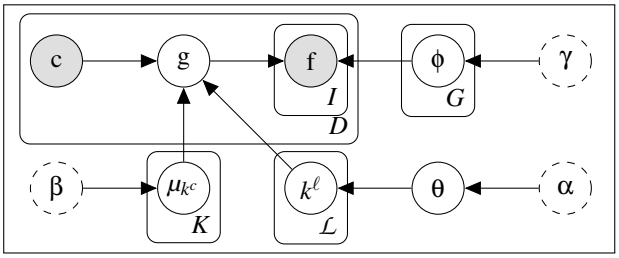

Figure 2: The plate diagram of the BCF model. Shaded nodes indicate observed variables, and dotted nodes indicate hyperparameters.

differ in their strength of association with feature types (e.g., the feature type `function` will be highly associated with TOOLS but less so with ANIMALS). BCF jointly optimizes categories and their featural representation: the learning objective is to obtain a set of meaningful categories, each characterized by relevant and coherent feature types.

The generative story and plate diagram for the BCF model are shown in Figures 1 and 2, respectively. The input to the model is a collection of *stimuli* $d \in \{1..D\}$ extracted from a large text corpus. Each stimulus consists of a target concept $c \in \{1..\mathcal{L}\}$ and its context $\mathbf{f} \in \{1..F\}$. We adopt a simple representation of context as the set of words making up the sentence c occurs in (except c). The model assigns concepts to categories $k \in \{1..K\}$ and features to feature types $g \in \{1..G\}$. It learns a set of concept clusters (i.e., categories), as well as a clustering over features (i.e., feature types), and a distribution over those feature clusters for each category (i.e., category-feature type associations). Specifically, the occurrences of a concept will be assigned a category, based on how similar the concept's feature types are compared to the feature types of all other potential categories. Simultaneously, upon observing a stimulus (i.e., a concept in context), the model assigns the context to a particular feature type based on its probability under all potential feature types, and the prior probability of observing that feature type with the concept's assigned category.

More formally, we can describe the model through the generative story given in Figure 1. We assume a global multinomial distribution over categories $Mult(\theta)$, drawn from a symmetric Dirichlet distribution with hyperparameter α. For each category k, we assume an independent set of multinomial parameters over feature types μ_k, drawn from a symmetric Dirichlet distribution with hyperparameter β. For each concept type ℓ, we draw a category k^{ℓ} from $Mult(\theta)$. Finally, for each feature type g, we draw a multinomial distribution over features $Mult(\phi_g)$ from a symmetric Dirichlet distribution with hyperparameter γ. With these global assignments in place, we can generate stimuli d as follows: we first retrieve the category k^{c^d} of the observed concept c^d; we then generate a feature type g^d from the category's feature type distribution $Mult(\mu_{k^{c^d}})$; and finally, for each feature position i we generate feature $f_{d,i}$ from the feature type's distribution $Mult(\phi_{g^d})$. The joint probability of the model over latent categories, latent feature types, model parameters, and data can be factorized as:

$$P(g,f,\mu,\phi,\theta,k|c,\alpha,\beta,\gamma) = \quad (1)$$
$$P(\theta|\alpha)\prod_{\ell}P(k^{\ell}|\theta)\prod_{k}P(\mu_k|\beta)\prod_{g}P(\phi_g|\gamma)$$
$$\prod_{d}P(g^d|\mu_{k^{c^d}})\prod_{i}P(f^{d,i}|\phi_{g^d}).$$

Since we use conjugate priors throughout, we can integrate out the model parameters analytically, and perform inference only over the latent variables, namely the category and feature type labels associ-

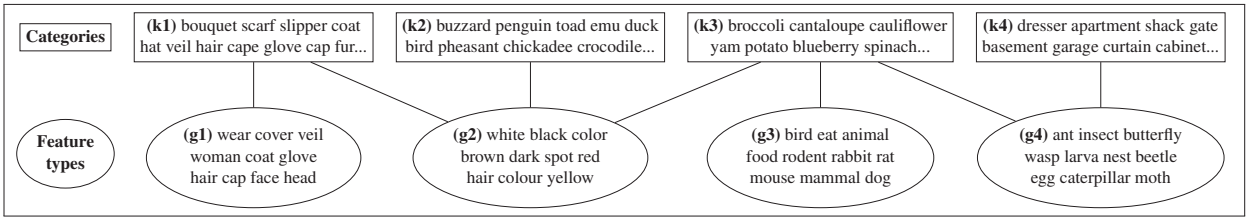

Figure 3: Example of categories (top) and feature types (bottom) inferred by the BCF model. Connecting lines indicate a strong association between the category and the respective feature type.

ated with the stimuli.

Exact inference in the BCF model is intractable, so we turn to approximate posterior inference to discover the assignments of latent variables that best explain our data. We construct a Gibbs sampler (Geman and Geman, 1984) which iteratively re-assigns single variables based on the current assignments of all other variables. One Gibbs iteration for our model consists of one sweep through the input stimuli, resampling feature type assignments from:

$$P(g^d_{k^{c^d}} = i | \mathbf{g}^{-d}_{k^{c^d}}, \mathbf{f}^{-}, k^{c^d}, \beta, \gamma) \quad (2)$$
$$\propto P(g^d_{k^{c^d}} = i | \mathbf{g}^{-d}_{k^{c^d}}, k^{c^d}, \beta) \times P(f^d | \mathbf{f}^{-}, g^d_{k^{c^d}} = i, \gamma),$$

followed by one sweep through the concept types, resampling category assignments from:

$$P(k^\ell = j | \mathbf{g}_{k^\ell}, \mathbf{k}^{-}, \alpha, \beta) \quad (3)$$
$$\propto P(k^\ell = j | \mathbf{k}^{-}, \alpha) \times P(\mathbf{g}_{k^\ell} | \mathbf{g}^{-}_{k^\ell}, k^\ell = j, \beta),$$

where $g^d_{k^{c^d}}$ denotes the feature type assignment to stimulus d given the category k^{c^d} of d's observed target concept c^d. k^ℓ refers to the category assignment of concept type ℓ, $\mathbf{g}_{k^\ell}$ refers to the feature type associations of category k^ℓ, and f^d refers to the observed features in stimulus d. The superscript $^-$ indicates the absence of the variable assignment(s) which are currently resampled from the current representation of the model state.

Figure 3 illustrates example output produced by our model, in terms of learnt categories, learnt feature types and their associations. Connecting lines indicate category-feature type associations. Feature types are shared across categories, e.g., categories CLOTHING (k1), BIRDS (k2), and FOOD (k3) are all associated with feature type `color` (g2).

4 Experimental Design

In this section we outline our experimental set-up for assessing the performance of the BCF model described above. We present our data set, briefly introduce the models used for comparison with our approach, and explain how system output was evaluated. We then report results on a series of experiments which evaluate the quality of the categories and feature types learnt by BCF.

Data Our experiments used basic-level target concepts (e.g., *cat* or *chair*) from two norming studies (McRae et al., 2005; Vinson and Vigliocco, 2008). In these studies, humans were presented with concepts and asked for each concept to produce a set of characteristic features. In a subsequent study (Fountain and Lapata, 2010), the concepts were classified into 41 categories (with possible multi-category membership), 34 of which we use as a goldstandard in our categorization experiments (comprising 492 concepts in total). We excluded very general categories such as THING or STRUCTURE, based on the intuition that it is difficult to identify characteristic features for them. As a heuristic concepts were excluded if they were close to the root of WordNet (e.g., with depth 2 or 4).

To obtain the input stimuli for the BCF model, we used a subset of the Wackypedia corpus (Baroni et al., 2009), an automatically extracted and POS tagged dump of the English Wikipedia. For each target concept, we identified one corresponding article in Wackypedia. Next, we extracted a set of stimuli which consists of (a) every sentence from the concept's corresponding article, and (b) any sentence in a different article which mentions the concept. This resulted in a data set of 63,076 stimuli which we split into 60% training, 20% development and 20% test.

We removed stopwords as well as words with a part of speech other than noun, verb, and adjective. Furthermore, we discarded words with an age of acquisition above 10 years (Kuperman et al., 2012) to restrict the vocabulary to frequent and generally familiar words.

Models and Parameters We compared the performance of BCF against BayesCat, a Bayesian model of category acquisition (Frermann and Lapata, 2014) and Strudel, a pattern-based model which extracts concept features from text (Baroni et al., 2010).

BayesCat induces categories, which are represented through a distribution over target concepts, and a distribution over features (i.e., individual context words). In contrast to BCF, it does not learn types of features. In addition, while BCF induces a hard assignment of concepts to categories, BayesCat learns soft distributions over target concepts for each category. Soft assignments can be converted into hard assignments by assigning each concept to its most probable category. We ran BayesCat on the same input stimuli as BCF, with the following parameters: the number of categories was set to $K = 40$, and the hyperparameters to $\alpha = 0.7, \beta = 0.1, \gamma = 0.1$. For the BCF model, we used the same number of categories, namely $K = 40$. The number of feature types was set to $G = 75$, and the hyperparameters to $\alpha = 0.5, \beta = 0.5$, and $\gamma = 0.1$. Parameters were tuned on the development set. For both models, we report results averaged over 10 Gibbs runs, each time we ran the sampler for 1,000 iterations. We used annealing during learning which proved effective for avoiding local optima.

Strudel automatically extracts features for concepts from text collections following a pattern-based approach. It takes as input a set of target concepts and a set of patterns, and extracts a list of features for each concept, where each concept-feature pair is weighted with a log-likelihood ratio expressing the pair's strength of association. Baroni et al. (2010) show that the learnt representations can be used as a basis for various tasks such as typicality rating, categorization, or clustering of features into types. In our experiments we obtained Strudel representations from the same Wackypedia corpus used for extracting the input stimuli for BCF (and BayesCat). Note that Strudel, unlike the two Bayesian models, is not a cognitively motivated *acquisition* model, but an optimized system developed with the aim of obtaining the best possible features from data.

4.1 Experiment 1: Evaluation of Categories

In our first experiment we evaluate the quality of the categories induced by the three models presented above. The models produce hard categorizations, however, the cognitive gold standard we use for evaluation (Fountain and Lapata, 2010) represents soft categories. We obtained a hard categorization by assigning members of multiple categories to their most typical category (typicality scores are provided with the data).[1]

Method BCF and BayesCat learn a set of categories which we can directly compare to the gold standard. For Strudel, we produce a categorization as follows: we represent each concept as a vector over features (obtained from Wackypedia), where each component corresponds to the concept-feature log-likelihood ratios provided by Strudel; following Baroni et al. (2010), we then cluster the vectors using K-means and the Cluto toolkit.[2] As for the other models, we set the number of categories to $K = 40$.

Metrics To assess the quality of the clusters produced by the models, we measure purity (*pur*; the extent to which each learnt cluster corresponds to a single gold class) as well as its inverse, collocation (*col*; the extent to which all items of a particular gold class are represented in a single learnt cluster). Both measures are based on set-overlap, and we also report their harmonic mean ($f1$; Lang and Lapata 2011). In addition, we report the V-measure ($v1$; Rosenberg and Hirschberg 2007) and its factors measuring the homogeneity of clusters (*hom*) and their completeness (*com*). The two factors intuitively correspond to purity and collocation, but are based on information-theoretic measures.

Results Our results are summarized in Table 1. They show that BCF and Strudel perform almost identically, and both outperform BayesCat. BCF *learns* the categories from data, whereas for Strudel

[1] http://homepages.inf.ed.ac.uk/s0897549/data/.

[2] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

	hom	*com*	*v*1	*pur*	*col*	*f*1
BCF	0.68	0.64	**0.66**	0.59	0.52	**0.55**
BayesCat	0.65	0.59	0.62	0.57	0.45	0.50
Strudel	0.70	0.62	**0.66**	0.61	0.48	0.54

Table 1: Model performance on the category induction task.

we construct the categories post-hoc after a highly informed feature extraction process (relying on grammatical patterns). It is therefore not surprising that Strudel performs well, and it is encouraging to see that BCF does too. Also, note that Strudel tends to learn very clean clusters at the cost of recall, whereas the tradeoff is less extreme for BCF. Again, this is expected given Strudel's pattern-based approach. While BCF and Strudel are constrained to assign each concept to only one category, BayesCat induces a soft categorization which is turned into a hard categorization in a post-learning step. While this setting allows for more flexibility, it also induces more uncertainty and results in categorizations which resemble the gold standard less closely compared to the two other models.

4.2 Experiment 2: Evaluation of Features

We next investigate the quality of the features our model learns. We do this by letting the model predict the right concept solely from a set of features. If the model has acquired informative features, they will be predictive of the unknown concept. Specifically, the model is presented with a set of previously unseen test stimuli with the target concept removed. For each stimulus, the model ranks all possible target concepts based on the features **f** (i.e., context words).

Method In our experiments we compared the ranking performance of BCF, BayesCat, and Strudel. For the Bayesian models, we directly exploit the learnt distributions. For BCF, we compute the score of a target concept c given a set of features as:

$$Score(c|\mathbf{f}) = \sum_{g} P(g|c)P(\mathbf{f}|g). \quad (4)$$

		pr@1	*pr*@10	*pr*@20	*avg*
BCF	full	**0.12**	**0.50**	0.63	56.1
	−tgt	0.09	0.40	0.53	78.5
BayesCat	full	0.11	0.49	**0.64**	**37.7**
	−tgt	0.09	0.39	0.53	52.4
Strudel	full	0.07	0.33	0.47	64.4
	−tgt	0.07	0.35	0.49	62.2

Table 2: Model performance on the concept prediction task. Precision at rank 1, 10, 20, and average rank assigned (*avg*). −tgt refers to the condition where we remove context words which are identical to the target concept as opposed to using the full context.

Similarly, for BayesCat we compute the score of a concept c given a set of features as follows:

$$Score(c|\mathbf{f}) = \sum_{k} P(c|k)P(\mathbf{f}|k). \quad (5)$$

For Strudel, we rank concepts according to the cumulative log-likelihood ratio-based association score over all observed features for a particular concept c:

$$Score(c|\mathbf{f}) = \sum_{f \in \mathbf{f}} association(c, f). \quad (6)$$

Metrics Since we can directly compare model predictions against the actual target concept of the stimulus, we report precision at rank 1, 10, and 20. We also report the average rank assigned to the correct concept. All results are based on a random test set of 2,000 previously unseen stimuli. To control for the possibility that the models are learning a strong (yet trivial) correlation between target concepts and identical words occurring as features, we also report results on a modification of our test set where we remove any mention of the target concept from the context, if present (the −tgt condition).

Results Our results on the concept prediction task are shown in Table 2. The Bayesian models outperform Strudel across all metrics and conditions. Strudel's extraction algorithm, which relies on predefined patterns, might be too restrictive with respect to the set of features it extracts and as a result they are not discriminative. BayesCat and BCF

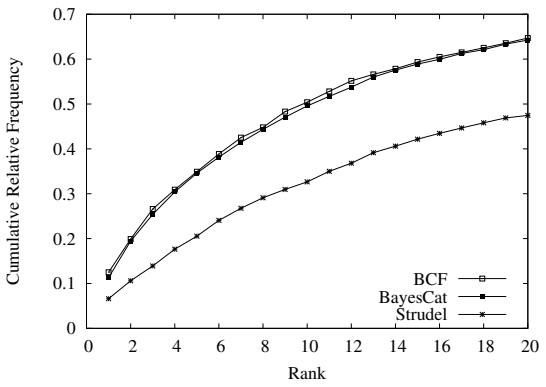

Figure 4: Number of times the correct target concept was placed within the top 20 ranks by BCF, BayesCat, and Strudel.

salmon	`journey move hundred mile strong current reproduce`				
BCF	**salmon**	tuna	goldfish	lobster	fish
BayesCat	fish	radio	goldfish	**salmon**	clock
Strudel	train	house	apartment	ship	car

finger	`avoid cut quick claw tip painful`				
BCF	tent	ski	peg	curtain	hut
BayesCat	eye	ear	spider	leg	hair
Strudel	**finger**	toe	hair	tail	hand

Table 3: Model output on the concept prediction task for *salmon* (top) and *finger* (bottom): the top part of each table shows the true concept (left) and the context provided to the model as input (right). The bottom part of the table shows the five most highly ranked concepts (left to right) for each model.

perform comparably given that they learn from exactly the same data and exploit local co-occurrence relations in similar ways. BayesCat produces better average rank scores than BCF, while achieving lower precision scores. This can be explained by the fact that BCF assigns low ranks to correct concepts more reliably than BayesCat. Figure 4 shows the relative cumulative frequencies of the ranks assigned by the three models. We display the top ranks 1 through 20 (out of 492). As can be seen, BCF performs slightly better than BayesCat. Pairwise differences between the systems are all statistically significant ($p \ll 0.01$); using a one-way ANOVA with post-hoc Tukey HSD test).

Note that performance decreases for the Bayesian models in the $-$tgt condition, i.e., when occurrences of the target concept are removed from the context. Strudel is less affected by this given its pattern-based learning mechanism which is not prone to associating target word types with themselves. However, repetitions are a natural phenomenon from a cognitive standpoint and it seems reasonable to consider multiple occurrences of a concept as a canonical feature of the learning environment.

Overall, the precision scores may seem low. However, the models rank a set of 492 target concepts; a random baseline would achieve a $pr@1$ of only 0.002%. In addition, the target concepts we are considering are by design highly confusable: they were selected so that they form categories and are thus bound to share some features which makes the prediction task harder. Example output for all three models is shown in Table 3. The models take context features "*journey move hundred mile strong*" and "*avoid cut quick claw tip*" as input and are expected to predict *salmon* and *finger*, respectively. Unlike Strudel, BCF and BayesCat rank *salmon* almost correctly and the other high ranked concepts are reasonable in the given context as well. For the second example, only Strudel predicts the correct concept correctly, but again the top-ranked concepts of the other two models are reasonable in the given context.

4.3 Experiment 3: Evaluation of Feature Types

In this suite of experiments we evaluate two aspects of the feature types induced by our model: (1) Are they *relevant* to their associated category? and (2) Do they form a *coherent* class? Our evaluation followed the intrusion paradigm originally introduced to assess the output of topic models (Chang et al., 2009). We performed two intrusion studies using Amazon's Mechanical Turk crowd-sourcing platform.

In the feature intrusion study, participants were shown examples of categories and their feature types both of which were represented as word clusters (see Figure 6 top). They were asked to detect the feature type which did not belong to the category. If a model creates *relevant* feature types, we would expect participants to be able to identify the intruder relatively easily. We also conducted a word intrusion

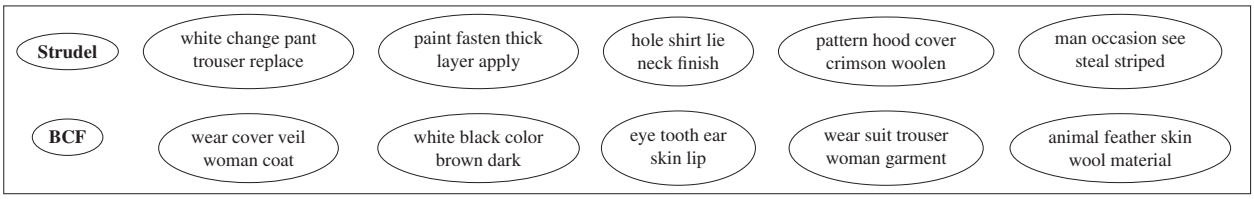

Figure 5: Example feature types learnt for the category CLOTHING by Strudel (top) and BCF (bottom).

'Select intruder feature type (right) wrt category (left).'	
ant hornet	○ egg female food young bird
moth flea	○ ant insect butterfly wasp larva
beetle wasp	● wear cover veil woman coat
cockroach	○ body air fish blood muscle

'Select the intruder word.'				
○	○	●	○	○
egg	female	box	young	bird

Figure 6: Illustration of the feature type intrusion task (top); and the word intrusion task (bottom).

study, where participants were shown a single feature type (again represented as a word cluster) and asked to detect the intruder feature/word (see Figure 6 bottom). If the features are overall *coherent* and meaningful, it should be relatively straightforward to identify the intruder.

Method We compared the feature types learnt by BCF and Strudel. We omitted BayesCat from this evaluation as it does not naturally produce feature types, rather it associates unstructured lists of features with categories. As mentioned earlier, Strudel does not induce feature types either, however, it associates concepts with features which can be post-processed to obtain feature types as follows. Given a category induced by Strudel (as explained in Experiment 1), we collected the features associated with at least half of the concepts in the category with a log likelihood score no less than 19.51.[3] We then clustered these features with K-means (using the Cluto toolkit) into $K = 5$ feature types.

For BCF, for each category k, we select the five feature types g with highest association $P(g|k)$, together with one intruder feature type g' which is highly associated with some other category k' but not with k. For Strudel we took the five feature types elicited through the procedure described above, and one random feature type from the global set of feature types. Each feature type was represented by a cluster of five words.

With respect to the word intrusion task, participants were only shown feature types (i.e., word clusters) irrespectively of the associated category. BCF feature types g were represented as the set of the five words w with highest probability $P(f|g)$. In addition, we added one intruder word which had low probability under g but high probability under some other feature type. For Strudel, we represented feature types as a random subset of five words, and added an additional intruder word from the global set of features.

For the feature type intrusion task, We evaluated a total of 40 categories for each model. Each participant assessed 10 categories per session (5 per model). Categories and feature types were presented in random order. For the word intrusion task, we evaluated a total of 66 feature types for each model. Participants saw 11 feature types per session, in randomized order. In both cases, we collected 10 responses per item.

Metrics We evaluated feature type relevance and coherence by measuring precision (the proportion of intruders identified correctly). We also use the Kappa coefficient to measure inter-subject agreement (Fleiss, 1981) on our two tasks.

Results Our results are presented in Table 4. Participants identify the intruder feature type correctly more than 50% of the time. The performance of Strudel is slightly better compared to BCF, both in terms of accuracy and Kappa (however the dif-

[3]Following Baroni et al. (2010), this number corresponds to a probability of co-occurrence below 0.00001, assuming independence.

	Feat Type Intrusion		Word Intrusion	
	Prec	Kappa	Prec	Kappa
BCF	0.52	0.23	**0.78**	**0.60**
Strudel	**0.56**	**0.26**	0.36	0.21

Table 4: Performance of Strudel and BCF on the feature type and word intrusion tasks. We report precision (Prec) and inter-subject agreement (Fleiss' Kappa; all Kappa values are statistically significant at $p \ll 0.05$).

ferences are not statistically significant, using a *t*-test). Again this is not surprising considering that Strudel's feature types were elicited through a highly informed, pipelined process. The results show that the simpler and cognitively plausible BCF model learns feature types of a quality comparable to a highly engineered, competitive system. Examples of feature types discovered by BCF and Strudel are shown in Figure 5, for the category CLOTHING. As can be seen, Strudel obtains a large number of action-related features (e.g., *replace, change, steal*). BCF creates more varied feature types. For example, the second cluster refers to external properties (e.g., color), and the last cluster contains CLOTHING materials.

Concerning the word intrusion task, we observe that participants are able to detect the intruder more accurately when presented with BCF feature types as compared to Strudel feature types (differences between Strudel and BCF are statistically significant at $p \ll 0.05$, again using a *t*-test). The results suggest that the feature types learnt by BCF are more coherent, and indeed express meaningful properties shared by concepts belonging to the same category. While being relevant to the category, Strudel's feature types do not seem to exhibit internal coherence to a similar extent. The mutual dependence of category formation and feature learning allows BCF to learn feature types which are both relevant and individually interpretable.

5 Discussion

In this paper we presented a cognitively motivated Bayesian model which jointly learns categories and their features, arguing that the two tasks are co-dependent. Our model learns from raw text without relying on elaborate post-processing and high-precision patterns. Evaluation of the inferred categories and their features shows that BCF performs competitively compared to a system specifically engineered to extract high quality features, despite the more complex learning objective, and the knowledge-lean approach. We approximate the cognitive learning environment with large text corpora. However, we do not claim to learn features qualitatively similar to features produced in human elicitation studies. Instead, we show, through a crowdsourcing-based human evaluation, that the learnt features are meaningful in that they are relevant to their associated category and form a coherent class.

An interesting direction for future work would be to learn feature types from multiple modalities (not only text) and to investigate how different information sources (e.g., visual or pragmatic input) influence feature learning. The BCF model learns descriptive feature types represented as a collection of feature values. In addition to such descriptive features (e.g., `behavior`) categories also possess *defining* features (e.g., `animate`) which are bound to one particular value. Extending the model in a way that allows to learn qualitatively different types of features is desirable from a cognitive perspective. We will also develop an incremental learning algorithm for joint category and feature learning (e.g., using sequential Monte Carlo methods such as Particle Filtering). In addition, it would be interesting to investigate the emergence of feature types with nonparametric Bayesian methods.

Finally, the BCF model can be applied to tasks beyond those discussed here. For example, one could learn definitions (aka features) of terms (aka concepts) in specialist fields (e.g., finance, law, medicine) or monitor how the meaning of words or concepts as represented by their features changes over time.

Acknowledgments We thank Micha Elsner and Charles Sutton for helpful discussions, William Schuler for his comments, and Carina Silberer for providing the Strudel features. We acknowledge the support of EPSRC through project grant EP/I037415/1.

References

Ahn, Woo-Kyoung. 1998. Why are different features central for natural kinds and artifacts?: the role of causal status in determining feature centrality. *Cognition* 69:135.

Anderson, John R. 1991. The adaptive nature of human categorization. *Psychological Review* 98:409–429.

Austerweil, Joseph L. and Thomas L. Griffiths. 2013. A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review* 120(4):817–851.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3):209–226.

Baroni, Marco, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science* 34(2):222–254.

Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*. pages 288–296.

Corter, James E. and Mark A. Gluck. 1992. Explaining basic categories - feature predictability and information. *Psychological Bulletin* 111(2):291–303.

Devereux, Barry, Nicholas Pilkington, Therry Poibeau, and Anna Korhonen. 2009. Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language & Computation* 7(2-4):137–170.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. Wiley, New York.

Fountain, Trevor and Mirella Lapata. 2010. Meaning representation in natural language categorization. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Portland, Oregon, pages 1916–1921.

Fountain, Trevor and Mirella Lapata. 2011. Incremental models of natural language category acquisition. In *Proceedings of the 33nd Annual Conference of the Cognitive Science Society*. Boston, Massachusetts, pages 255–260.

Frermann, Lea and Mirella Lapata. 2014. Incremental Bayesian learning of semantic categories. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*. pages 249–258.

Geman, Stuart and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6):721–741.

Goldstone, Robert L., Yvonne Lippa, and Richard M. Shiffrin. 2001. Altering object representations through category learning. *Cognition* 78:27–43.

Kelly, Colin, Barry Devereux, and Anna Korhonen. 2014. Automatic extraction of property norm-like data from large text corpora. *Cognitive Science* 38(4):638–682.

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods* 44(4):978–990.

Lang, Joel and Mirella Lapata. 2011. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK., pages 1320–1331.

McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods* 37(4):547–59.

Nosofsky, Robert M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14:700–708.

Reed, Stephen K. 1972. Pattern recognition and categorization. *Cognitive psychology* 3(3):382–407.

Riordan, Brian and Michael N. Jones. 2011. Redundancy in perceptual and linguistic experience:

Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science* 3(2):303–345.

Rosenberg, Andrew and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 410–420.

Sanborn, Adam N., Thomas L. Griffiths, and Daniel J. Navarro. 2006. A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada, pages 726–731.

Schyns, Philippe G and Luc Rodet. 1997. Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23:681–696.

Spalding, T. L. and B. H. Ross. 2000. Concept learning and feature interpretation. *Memory & Cognition* 28:439–451.

Steyvers, Mark. 2010. Combining feature norms and text data with topic models. *Acta Psychologica* 133(3):234–243.

Vinson, David and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods* 40(1):183–190.